Google Research Blog

The latest news from Research at Google

Teaching machines to read between the lines (and a new corpus with entity salience annotations)

Posted: Monday, August 25, 2014

Posted by Dan Gillick,
Research Scientist, and Dave Orr, Product Manager

Language understanding systems are largely trained on freely available data, such as the Penn Treebank, perhaps the most widely used linguistic resource ever created. We have previously released lots of linguistic data ourselves, to contribute to the language understanding community as well as encourage further research into these areas.

Now, we're releasing a new dataset, based on another great resource: the New York Times Annotated Corpus, a set of 1.8 million articles spanning 20 years. 600,000 articles in the NYTimes Corpus have hand-written summaries, and more than 1.5 million of them are tagged with people, places, and organizations mentioned in the article. The Times encourages use of the metadata for all kinds of things, and has set up a forum to discuss related research.

We recently used this corpus to study a topic called "entity salience". To understand salience, consider: how do you know what a news article or a web page is about? Reading comes pretty easily to people -- we can quickly identify the places or things or people most central to a piece of text. But how might we teach a machine to perform this same task? This problem is a key step towards being able to read and understand an article.

One way to approach the problem is to look for words that appear more often than their ordinary rates. For example, if you see the word “coach” 5 times in a 581 word article, and compare that to the usual frequency of “coach” -- more like 5 in 330,000 words -- you have reason to suspect the article has something to do with coaching. The term “basketball” is even more extreme, appearing 150,000 times more often than usual. This is the idea of the famous TFIDF, long used to index web pages.

Congratulations to Becky Hammon, first female NBA coach! Image via Wikipedia.

Term ratios are a start, but we can do better. Search indexing these days is much more involved, using for example the

distances between pairs of words on a page to capture their relatedness. Now, with the Knowledge Graph, we are beginning to think in terms of entities and relations rather than keywords. “Basketball” is more than a string of characters; it is a reference to something in the real word which we already already know quite a bit about.

Background information about entities ought to help us decide which of them are most salient. After all, an article’s author assumes her readers have some general understanding of the world, and probably a bit about sports too. Using background knowledge, we might be able to infer that the WNBA is a salient entity in the Becky Hammon article even though it only appears once.

To encourage research on leveraging background information, we are releasing a large dataset of annotations to accompany the New York Times Annotated Corpus, including resolved Freebase entity IDs and labels indicating which entities are salient. The salience annotations are determined by automatically aligning entities in the document with entities in accompanying human-written abstracts. Details of the salience annotations and some baseline results are described in our recent paper: A New Entity Salience Task with Millions of Training Examples (Jesse Dunietz and Dan Gillick).

Since our entity resolver works better for named entities like WNBA than for nominals like “coach” (this is the notoriously difficult word sense disambiguation problem, which we’ve previously touched on), the annotations are limited to names.

Below is sample output for a document. The first line contains the NYT document ID and the headline; each subsequent line includes an entity index, an indicator for salience, the mention count for this entity in the document as determined by our coreference system, the text of the first mention of the entity, the byte offsets (start and end) for the first mention of the entity, and the resolved Freebase MID.

```
1453526     Should Tyco's Auditors Have Told More?
0     1     18    Tyco International 20    38    /m/02b5ky
1     0     2     David Boies 412   423   /m/09p37
2     1     2     PricewaterhouseCoopers    536   558   /m/012_78
3     0     2     ADT   669   672   /m/04q5hl
4     0     5     Securities and Exchange Commission    1260  1294  /m/0fbr4
5     0     1     Manhattan   1303  1312  /m/0cc56
6     0     2     L. Dennis Kozlowski     2157  2176  /m/06cs61
```

Features like mention count and document positioning give

reasonable salience predictions. But because they only describe what's explicitly in the document, we expect a system that uses background information to expose what's implicit could give better results.

Download the data directly from Google Drive, or visit the project home page with more information at our Google Code site. We look forward to seeing what you come up with!

Labels: datasets, Entity Salience, Natural Language Processing
82 comments

Google Research Awards: Summer 2014

Posted: Wednesday, August 20, 2014

posted by
Maggie
Johnson, Director of Education and University Relations

We have just completed another round of the Google Research Awards, our biannual open call for proposals on computer science-related topics including systems, machine perception, structured data, robotics, and mobile. Our grants cover tuition for a graduate student and provide both faculty and students the opportunity to work directly with Google researchers and engineers.

This round we received 722 proposals, an increase of 5% over last round, covering 44 countries on 6 continents. After expert reviews and committee discussions, we decided to fund 110 projects. The subject areas that received the highest level of support were systems, human-computer interaction, mobile, and machine perception, with 22% of the funding awarded to universities outside the U.S.

We introduced three new topics this round, representing important new research areas for Google. Computational neuroscience looks at the information processing properties of the brain and nervous system. One funded proposal will study scene recognition in this context. A second new area is physical interactions with devices. With the introduction of new paradigms such as Google Glass, we can study how such devices expand our processing capabilities. The third new area is online learning at scale, which covers topics such as teacher-

student interaction at scale, data-driven adaptive learning, and innovative assessment methods.

Congratulations to the well-deserving recipients of this round's awards. If you are interested in applying for the next round (deadline is October 15), please visit our website for more information.

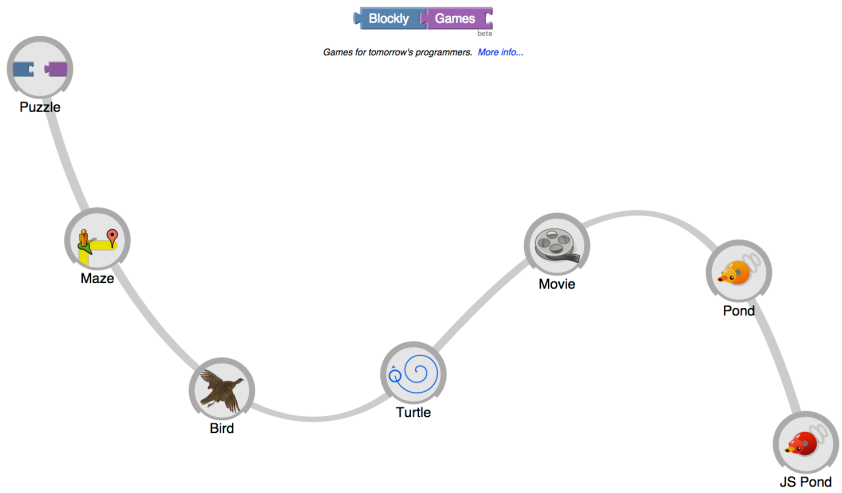
Labels: grants, Research Awards, University Relations
19 comments

Summer Games: Learn to Program

Posted: Monday, August 11, 2014

Posted by
Jennifer Vaden
Barth, Executive Assistant

Looking for ways to engage your kids in constructive, meaningful learning? We've just launched Blockly Games, our next extension of Blockly, a web-based graphical programming environment. As part of the generation of new programming environments that provide a more accessible introduction to coding, Blockly Games allows users to create and run programs by arranging blocks with a simple click, drag and drop.



Blockly Games requires little or no typing, which facilitates young or novice programmers to learn core coding principles in an intuitive way. By minimizing the use of syntax, users are able to focus on the logic and concepts used by computer scientists, progressing at their own pace as they venture through mazes

and more advanced arenas.

Blockly was featured during the 2013 Computer Science Education week where people of all ages tried programming for the first time. Blockly is universally accessible with translations for a number of languages, including German, Vietnamese, Russian and even Klingon.

We encourage you and your child to explore Blockly Games, where novice programmers of any age begin to learn together. With Blockly Games, the whole family can learn and master basic computer science concepts.

Labels: Education, K-12
132 comments

Doing Data Science with coLaboratory

Posted: Friday, August 08, 2014

Posted by
Kayur Patel,
Kester Tong, Mark Sandler, and Corinna Cortes, Google Research

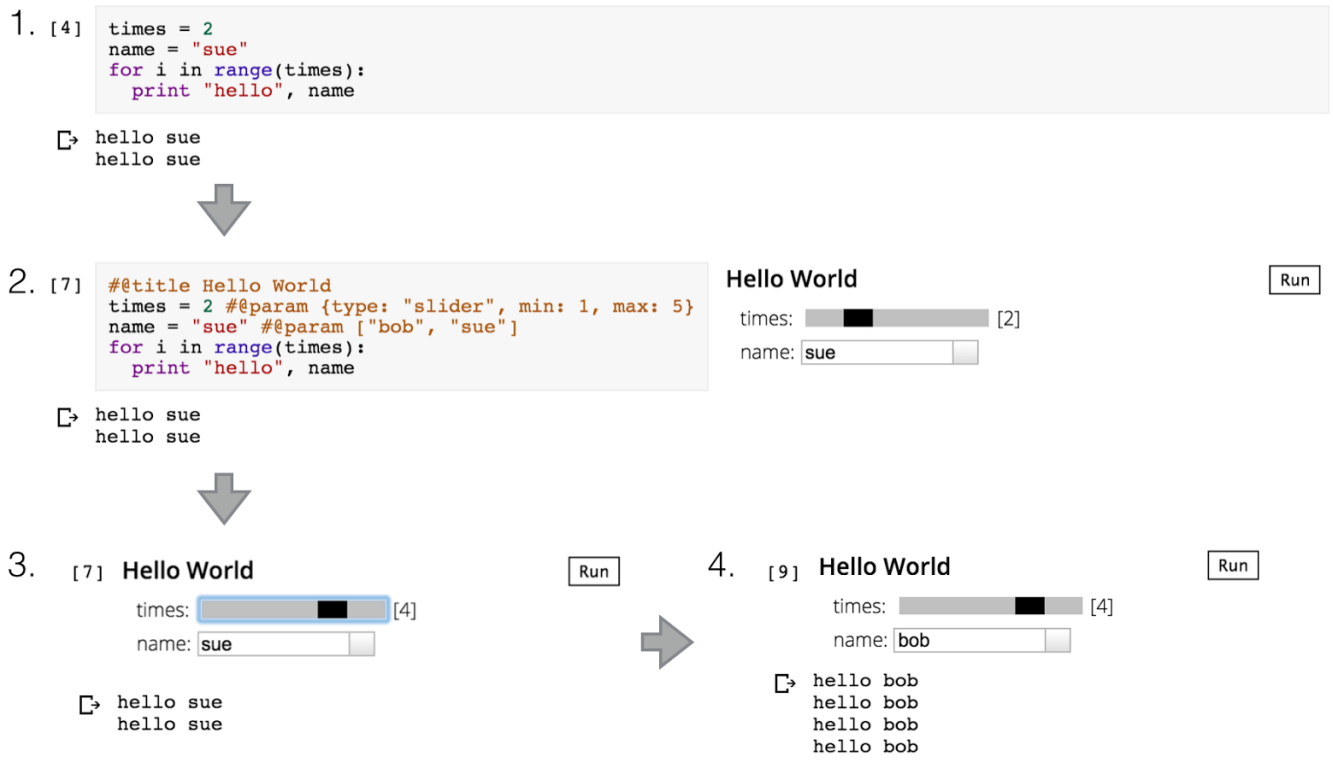
Building products and making decisions based on data is at the core of what we do at Google. Increasingly common among fields such as journalism and government, this data-driven mindset is changing the way traditionally non-technical organizations do work. In order to bring this approach to even more fields, Google Research is excited to be a partner in the coLaboratory project, a new tool for data science and analysis, designed to make collaborating on data easier.

Created by Google Research, Matthew Turk (creator of the yt visualization package), and the IPython/Jupyter development team, coLaboratory merges successful open source products with Google technologies, enabling multiple people to collaborate directly through simultaneous access and analysis of data. This provides a big improvement over ad-hoc workflows involving emailing documents back and forth.

Setting up an environment for collaborative data analysis can be a hurdle, as requirements vary among different machines and operating systems, and installation errors can be cryptic. The

coLaboratory Chrome App addresses this hurdle. One-click installs coLaboratory, IPython, and a large set of popular scientific python libraries (with more on the way). Furthermore, because we use Portable Native Client (PNaCl), coLaboratory runs at native speeds and is secure, allowing new users to start working with IPython faster than ever.

In addition to ease of installation, coLaboratory enables collaboration between people with different skill sets. One example of this would be interactions between programmers who write complex logic in code and non-programmers who are more familiar with GUIs. As shown below, a programmer writes code (step 1) and then annotates that code with simple markup to create an interactive form (step 2). The programmer can then hide the complexity of code to show only the form (step 3), which allows a non-programmer to re-run the code by changing the slider and dropdowns in the form (step 4). This interaction allows programmers to write complex logic in code and allows non-programmers to manipulate that logic through simple GUI hooks.



For more information about this project please see our talks on collaborative data science and zero dependency python. In addition to our external partners in the coLaboratory project, we would like to thank everyone at Google who contributed: the Chromium Native Client team, the Google Drive team, the Open Source team, and the Security team.

Labels: data science, IPython, open source

76 comments

Newer Posts Home Older Posts